Computers and the Humanities **32:** 153–184, 1998.

Compatibility in Interpretation of Relations in EuroWordNet

PIEK VOSSEN[1,2], LAURA BLOKSMA[1,3], ANTONIETTA ALONGE[4,5], ELISABETTA MARINAI[4,6], CAROL PETERS[7], IRENE CASTELLON[8,9], ANTONIA MARTI[8,10] and GERMAN RIGAU[11]
[1]*Universiteit van Amsterdam, Faculteit Geesteswetenschappen, Spuistraat 134, 1012 VB Amsterdam, The Netherlands; E-mail:* [2]*Piek.Vossen@hum.uva.nl,* [3]*lbloks@hum.uva.nl;* [4]*Istituto di Linguistica Computazionale, CNR, Via della Faggiola 32, 56100 Pisa, Italy; E-mail:* [5]*aalonge@pg.tecnonet.it,* [6]*elisabetta@ilc.pi.cnr.it;* [7]*Istituto di Elaborazione della Informazione, CNR, Via S. Maria, 46, 56126 Pisa, Italy; E-mail: carol@iei.pi.cnr.it;* [8]*Universitat de Barcelona, Departament de Filologia,Romanica Linguistica General, Gran Via 505, 08007 Barcelona, Spain; E-mail:* [9]*castellon@lingua.fil.ub.es,* [10]*amarti@lingua.fil.ub.es;* [11]*Universitat Politècnica de Catalunya. Jordi Girona Salgado, 1-3, 08034 Barcelona Spain; E-mail: g.rigau@lsi.upc.es*

Key words: overlapping relations and lexical gaps, sense differentiation

Abstract. This paper describes how the Euro WordNet project established a maximum level of consensus in the interpretation of relations, without loosing the possibility of encoding language-specific lexicalizations. Problematic cases arise due to the fact that each site re-used different resources and because the core vocabulary of the wordnets show complex properties. Many of these cases are discussed with respect to language internal and equivalence relations. Possible solutions are given in the form of additional criteria.

1. Introduction

The main objective of Euro WordNet is to build a multilingual database with wordnets for several languages. This multilingual database can be used directly in applications such as cross-language information retrieval or for comparison of the different wordnets. However, comparison and cross-linguistic retrieval only make sense when the separate wordnets are compatible in coverage and interpretation of relations. In (Rodriguez et al., this volume) it is described how we established compatibility in coverage of vocabulary. This paper deals with the compatibility in the interpretation of the relations.

We ensured a minimal level of consensus on the interpretation of lexical semantic relations by using explicit tests to verify the relations across words (as detailed in (Alonge et al., this volume)). This interpretation is in principle given by substitution tests (comparable to the diagnostic frames, (Cruse, 1986)) for each relation. Despite these tests it is nevertheless often difficult to decide on how the relations should be encoded. The tests do not always yield a clear intuition and in some cases there are still several possibilities open.

Especially the more fundamental and frequently used Base Concepts often turn out to be very complex. Typically, these Base Concepts have the following properties:

- They belong to high polysemous entries, having many and often vaguely-distinguished meanings (e.g. *make* which has 31 senses as a verb, *go* which has 28 senses as a verb, *head* which has 23 senses as a noun).
- They belong to large synsets; having more than average number of synonyms (e.g. *human body* 1 which has 14 synset members).
- They have poor definitions exhibiting circularity, co-ordination of genus words, void genus words.
- They have inconsistent patterns of hyponyms and hyperonyms across resources.
- They have a variety of syntactic properties.
- They are frequently used in daily language.

Still, these words make up the core of the wordnets, representing major semantic implications and clusters, which are carried over to the rest of the vocabulary. It is therefore extremely important that we still achieve a maximum of consensus on the encoding of these concepts across the sites, without loosing the possibilities to encode language-specific lexicalizations. For this we exchanged and compared specific problematic cases and had discussions on principles and strategies in order to deal with classes of problems.

This paper is a report on these discussions. We have given the solutions in the form of additional criteria, which can be used to make a decision, and by giving typical examples, which can be used for comparison. In Section 2 we discuss the problems with encoding the language-internal relations, especially with respect to our core vocabulary. In subsections we describe the typical problems that may arise, caused by differences in sense distinction, incompleteness and/or inconsistency in information and overlapping relations. In Section 3 we discuss the problems related to specifying the correct equivalence relations with the WordNet1.5 synsets, caused by lexical gaps, differences in sense distinction across wordnets and mismatches of senses.

It is important to note that the procedure outlined and the problems discussed are not typical for the encoding process. In most cases, the relations are obvious and the encoding is straightforward. In this document we focus on the problematic cases and describe the (possible) solutions we found to ensure maximum compatibility. Finally, we assume that the reader is familiar with the other papers in this volume.

2. Strategies for Encoding Language-internal Relations

In EuroWordNet we re-use existing Machine Readable Dictionaries and Lexical Databases as far as possible, which is more cost-effective than starting from scratch. Therefore the information in the resources serves as a starting point for encoding the semantic relations. The general approach towards defining the relations for a word meaning can be described as a set of steps:

1. determine the appropriate division for the relevant senses of a word
2. determine the synsets
3. determine the hyperonyms for a synset
4. determine the hyponyms for a synset
5. determine the near synonyms
6. determine the other relations relevant to the synset
7. determine the equivalence relations with the WordNet1.5 synsets

Obviously, the order of these steps is not mandatory. Each site builds their wordnet according to the scheme that fits best their resources and tools. In some cases, sites may arrive at step 1 after having worked on step 2 up to 4, and in other cases, they may start with the translation from WordNet1.5 (step 7). The Spanish group, for example, first translates the WN1.5 synsets into Spanish (step 7), next they create the Spanish synsets (step 2) and take over the hyponymy relations from WN1.5 (steps 3 and 4). After that, steps 5 and 6 are performed and if necessary step 1. The order in this document is only given as a rule of thumb for clarification purposes, it is by no means prescriptive.

In the next subsections we will discuss the problems that arise when determining the appropriate sense distinction (step 1). Next, we will look at the problem of deriving comprehensive and consistent patterns of relations for word meanings. Finally, we will discuss various border cases where the choice between the semantic relations appears less clear (steps 2–6). Step 7 is discussed in Section 3.

2.1. DIFFERENCES IN SENSE DISTINCTION

As already mentioned, all sites use the information in their resources as a starting point for building the wordnets. This means that the sense distinction made by the resources is in principle accepted and then verified. In most cases there is no reason to alter the distinction. However, in other cases, the differences are very subtle, which can lead to many closely related senses, or condensed to only a single sense (as discussed by Jacobs, 1991; Atkins and Levin, 1988). Here, we would like to discuss those cases that are problematic when building our wordnets. We distinguish between two types of problems:

- over-differentiation of senses
- under-differentiation of senses

2.1.1. *Over-differentiation of senses*

In the case of over-differentiation the motivation for distinguishing different senses is not clear or intuitions vary. In the following examples the definitions of the different senses are more or less similar.

(1.) a	*draaien* (to run)	1	functioneren (to function)
		2	aan de gang zijn (working)
b	*scuola* (school):	3a	attivit á rivolta a far apprendere una o piú discipline (activity aimed at causing to learn one of more disciplines);
		3b	l'insegnamento (teaching)
		3c	indirizzo di studio o metodo didattico e pedagogico adottato (line of study or didactical and pedagogical method adopted).

Although formulated in a different way the two senses of *draaien* (to run) in Dutch boil down to the same thing. Another example is represented by the Italian word *scuola* (school). In the main Italian source, there are 11 word-senses for this term, distributed variously over 5 principal word-meanings, of which a few distinctions are very subtle. In these cases it might be helpful to look at the rest of the information provided for the senses.

If a sense does not provide any really different information, we assume that there is an over-differentiation and one of the senses can be removed. This is the case of the Spanish entry *sopa* (soup):

(2.) *sopa*	1	Pedazo de pan empapado en cualquier líquido (A piece of bread soaked in any liquid)
	2	Plato compuesto de un líquido alimenticio y rebanadas de pan (Dish composed by a nutritive liquid and pieces of bread)
	3	Plato compuesto de rebanadas de pan, fécula, arroz, fideos, etc., y el caldo de la olla u otro análogo en que se han cocido (Dish composed of pieces of bread, starch, rice, noodle, etc. and stock . . .)

4 Pasta, fécula o verduras que se mezclan con el caldo en el plato de este mismo nombre

(Pasta, starch or vegatables mixed with the stock in the dish with the same name)

5 Comida que dan a los pobres en los conventos

(Meal served to the poor in a religious establishment in the convent)

6 Rebanadas de pan que se cortan para echarlas en el caldo

(Slices of bread cut and added into the stock)

Sense 6 is related to sense 1 by a hyponymy relation (where stock is a particular portion of "any liquid") both describing the main ingredients of the soup. This is also the case for sense 4 where, the ingredients added to the stock are different. On the other hand, sense 2 is included in sense 3 describing both the complete dish. Sense 5 is describing the same dish as sense 2 and 3 but is related to a particular situation. We can thus merge sense 1, 4 and 6 into a single meaning, and sense 2, 3 and 5 into another meaning. If the senses differ in any other kind of information, it is more difficult to make a decision. There are numerous reasons why a dictionary might split an entry into multiple senses, only some of which have to do with meaning (Gale et al., 1993). Often, senses are distinguished because of differences in morpho-syntactic properties:

- part-of-speech (nouns vs. adjectives, etc.).
- syntactic features (person, number, gender, etc.).
- valency structures (transitive vs. intransitive verbs, etc.).

The relevance of different grammatical and stylistic properties for distinguishing senses depends on the strictness of the definition of synonymy, where stylistic differences are usually not considered as differences of meaning. As a rule of thumb, we can state that morpho-syntactic properties that correlate with semantic differences, or with one of the semantic relations distinguished, should certainly be taken seriously. This is the case for many of the alternations of verbs (e.g. transitive/intransitive-causative/inchoative alternations, see (Levin, 1993) for an overview of English verbs):

(3.) a	*cambiare*	1	intransitive
	(to change)		to become different
		2	transitive
			to make different
	cambiare	2	causes cambiare 1

b	*bewegen*	1	intransitive
	(to move)		(to change place or position)
		2	transitive
			(to cause to change place or position)
		6	reflexive
			((of people, animals) to change place or position)
	bewegen	2	causes bewegen 1

Here we see that Italian *cambiare* 1 and 2 (change) exhibit a transitive/intransitive alternation which correlates with a difference in causation. Something similar holds for different senses of *bewegen* (move) in Dutch, which refer as intransitive verbs to a non-causative change-of-position and as transitives to the causation of such a change (this also holds for *mover* (move) in Spanish and *muovere* (move) in Italian).

Another typical example is given by countable/uncountable variation of nouns. For example, the uncountable Italian word *acqua* (water) signifies specific/specialized senses when it is used in the plural, such as: *acque territoriali* (coastal waters), *acque termali* (thermal waters), *acque minerali* (mineral waters). Another case is given by Dutch *zaad* (seed) which, as a countable noun, refers to *a single mature fertilized plant ovum* and as an uncountable noun to an amount of this. Clearly, the relation between these senses can be expressed by one of the semantic relations in EuroWordNet: *zaad 2* HAS_MERONYM *zaad 1*.

In other cases, differences in morpho-syntactic features do not carry any semantic distinction as, for example, the change of gender in the Italian word *zucchino* or *zucchina* which means the same vegetable and is used indifferently in both morpho-syntactic forms. Another typical example is formed by Dutch plural variants, such as *aardappels* (potatoes) and *aardappelen* (potatoes). There may be a difference in style but these are typically seen as variants of the same meaning. If such stylistic or formal properties are the only reason for making a distinction in different senses we follow the strategy of collapsing the senses and storing the variations as stylistic or formal variation of a single sense:[1]

(4.) Variant
key = aardappel
pos = NOUN
plural-form = aardappels; aardappelen
countable = true

In all cases, where there is still some doubt about the similarity or equivalence of different senses, either due to subtle differences in the information or examples, the senses can be connected by a NEAR_SYNONYM relation. In this way, we at least ensure that very close meanings are grouped together in contrast to other co-hyponyms (words that have the same hyperonym or class) which are clearly considered as distinct.

At times we find that two senses have very different definitions but can still be considered as cases of over-differentiation. Two specific situations are often encountered:

- pragmatic specialisation
- different conceptualisation

Pragmatic specialisation is the phenomenon where a general word is used as a variant to refer to a more specific concept: a *car* can also be referred to using *vehicle* or even *thing*. In some cases this usage has lead a lexicographer to distinguish a separate sense for the specific use of such a general word, e.g. in WordNet1.5:

(5.) mixture 1 (a substance consisting of two or more substances mixed together (not in fixed proportions and not with chemical bonding))

HAS_HYPERONYM

substance, matter (that which has mass and occupies space; "an atom is the smallest indivisible unit of matter")

mixture 2

HAS_HYPERONYM

foodstuff (a substance that can be used or prepared for use as food)

In this case, a hyponymy-relation holds between the specific sense of *mixture* used for food and the general sense of the word *mixture*. Whenever the specific sense is fully predictable the sense is strictly speaking superfluous. Predictability follows from the fact that no idiosyncratic properties are implied (no specialisation) and the principle can productively be applied to any other specific referent: *mixture* can also be used to refer to other substances with some function *paint*, *explosives*, *gases*. Predictable specialisations can be omitted (Roventini, 1993). This was clearly the case for the Spanish entry of *sopa* soup shown above, where sense 5 describes a specific pragmatic difference with respect senses 2 and 3 because it refers to the people who receives the soup and the place where the soup is served.

Another possibility is that the different senses reflect different perspectives or conceptualisations of the same thing. In Italian, for example, some pieces of cutlery or chinaware can both be seen as containers and as the quantity of food or drink contained. So we find this double sense for terms such as *cucchiaio* (spoon), *tazza* (cup), *bicchiere* (glass), *piatto* (plate), etc. Traditional dictionaries often do not allow for the expression of multiple perspectives and the traditional way of defining words does not promote this. This either results in the omission of one perspective (e.g. certain items of *cutlery* are either classified as a *quantity* or as a *container*) or in the separation in different senses. However, in EuroWordNet (and also in WordNet1.5), it is possible to have multiple hyperonyms reflecting these perspectives of the same concept or meaning (possibly by using disjunction or conjunction), as is illustrated by the WordNet1.5 solution for *spoon*:

(6.) *spoon* (a piece of cutlery with a shallow bowl-shaped container and a handle; used to stir or serve or take up food)

HAS_HYPERONYM

cutlery (implements for cutting and eating food)

container (something that holds things, especially for transport or storage)

The co-ordination-test (Zwicky and Sadock, 1975) shows that both conceptualisations can easily be combined, e.g. “It is a spoon therefore it is a piece of cutlery and a container”. In this case it is valid to merge the information of the two senses in a single sense, as is done for *spoon* in WordNet1.5. Especially when it turns out that multiple sources classify the same concept differently it may be possible to merge multiple senses in a particular source in which these different classifications are split.

2.1.2. *Under-differentiation of senses*

The opposite situation in which different senses are collapsed in a single definition also occurs frequently in dictionaries. Mostly this is done using co-ordination, e.g.:

(7.) a	*automatisering* 1	het automatisch *maken* of *worden*
	(automation)	(to *make* automatic or *become* automatic)
	beleefdheid 1	beleefde *handeling* of *uiting*
	(politeness)	(polite *act* or *utterance*)
	beroepsopleiding 1	*cursus* of *school*
	(occupational-training)	(*course* or *school*)
b	abombar 1	Dar o adquirir forma convexa [alguna cosa]
		(to give or to adopt [something] a convex shape)
	absorber 7	Retener o captar energía por medio de un material.
		(to *keep* or to attract energy by means of a material)
	achicharrar 1	*Freír*, *asar* o *tostar* [un manjar] hasta que tome sabor a quemado.
		(to *fry*, to *roast* or to *toast* [a food] until it takes a burned flavour)

For some of these examples it appears difficult to combine the hyperonyms of the definitions (underlined):

(8.) ∗de ene beroepsopleiding heeft een nieuw adres en de andere wordt twee keer gegeven.
(the one occupational-training has a new address and the other is given twice)

Since something cannot be an *institute* with an *address* and an *event* at the same time it seems to make more sense to distinguish two senses here. Furthermore, as separate senses it is possible to express the semantic relation between them; *beroepsleiding* 1 ROLE_LOCATION *opleiding 2*. In the case of verbs such as *maken* (make) and *worden* (become) we can state that they represent alternations of meanings which can be related using a CAUSES relation.[2]

Another pattern of co-ordination is illustrated by the following examples:

(9.) a *uitdaging* 2 *zaak*, *daad* of *uiting* die prikkelt tot een reactie
(challenge) (a *thing*, *act* or *utterance* which calls for a response)
toevlucht 1 *persoon*, *zaak*, *plaats* waar men bescherming zoekt.
(resort) (*person*, *thing*, *place* where one hopes to find protection)
b *antecedente* 2 a*cción*, *dicho* o *circunstancia* anterior, que sirve para juzgar hechos posteriores (previous *act*, *saying* or *circumstance*, which can be used to judge posterior events)
audición 2 Concierto, recital o lectura en público
(public *concert*, *recital* or *reading*)
batido 4 *Clara*s, *yema*s o *huevo*s batidos
(white, yolk or shake eggs)
bodrio 5 *Objeto*, *persona* o *actividad* desagradable o fea
(ugly *object*, *person* or *activity*)
c *disperazione* 2 cosa o persona che causa infelicità
(thing or person causing unhappiness)
problema 2 cosa o persona che causa problemi
(thing or person causing trouble)

Just as with the previous disjunctive hyperonyms we see that the test for distinguishing senses shows that the hyperonyms are incompatible:

(10.) ∗If it is a challenge then it is a thing and a person at the same time.

Strictly speaking, we should therefore split the sense into separate senses. However, how many senses do we have to distinguish here? The difference with the previous examples is that the range of entities is not restricted at all. There is an open range of referents for which some examples are listed: the list can easily be extended without changing the meaning. Conceptually, the test causes anomaly but in the case of the open denotation range the classifications do not motivate a separation of senses. Apparently, there is not one way to classify the referent, and the semantics of the word fully depends on the role or involvement it has with the event or situation expressed.

Since there may be an open range of entities it does not make much sense to split these in different senses. We therefore maintain a single sense for the definition where we can indicate the range of entities with disjunction of the hyperonym relation, but more important than the hyponymy-relation is the ROLE-relation with the predicate denoting the event:

(11.) *uitdaging* 2 (a challenge)

HAS_HYPERONYM disjunct:	*zaak* 1 (thing)
HAS_HYPERONYM disjunct:	*daad* 1 (deed)
HAS_HYPERONYM disjunct:	*uiting* 1 (utterance)
ROLE_AGENT:	*uitdagen* 1 (to challenge)

As long as the fundamental role relation is captured, the hyponymy relation may also be omitted.

The same problem also arises when no explicit genus term appears in definition. Consider for instance the following Spanish examples:

(12.) comida 1 lo que se come.
(*food, whatever that may be eaten*).
denunciante 1 que hace una denuncia.
(*informer, who makes a report*).

The genus words *lo que* and *que* are pronouns that hardly differentiate. There are 2,362 noun definitions (2%) in the Spanish monolingual resource with such void heads. Similar patterns have also been found in the resources for the other languages (Vossen, 1995). In the case of a void head or genus word the denotational range is not even specified and the role/involved relation is the only relation that can be used.

Obviously, it will not always be possible to distinguish cases where co-ordinated hyperonyms should be split for different senses or combined in a single sense. To some extent, the decision to split or merge senses depends on common practice.

2.2. COMPLETENESS AND CONSISTENCY OF INFORMATION

After establishing a good view on the different senses of a word, the next step is to identify all the relevant words that should be related to such a meaning. One of the challenges for building a consistent lexical database is perhaps not so much the quality of the data but more its incompleteness: i.e. what information is not given. It is an inherent property of our minds that we cannot easily recall all possible information and relevant meanings actively, but that we can very easily confirm information presented to us. Especially, when dealing with large coverage resources such as generic lexical databases it is impossible to predict the total potential of relations.

The general way of overcoming the problem of completeness is to combine information from different resources. It is for example possible to treat the definitions in different monolingual dictionaries as a corpus and to collect those definitions that have relevant co-occurrences of words. Following (Wilks et al., 1993) two words are co-occurrent if they appear in the same definition (word order in definitions is not taken into account). This method has been applied to a monolingual Spanish dictionary, from which a lexicon of 300,062 co-occurrence pairs for 40,193 word forms was derived (stop words were not taken into account). Table I, for example, shows the first eleven words (ordered by Association Ratio[3] score) out of 360 that co-occur with *vino* (wine). In this sample, we can see many implicit relations. Among others, hyponyms (*vino tinto*), hyperonyms (*licor or bebida)*, sisters (*mosto or jerez)*, inter-category relations (*beber*), places were the wine are maked/stored (cubas), fruit from which the wine is derived (*uva*), properties (*sabor*), etc. Such a raw list can be used as a starting point for establishing the construction of comprehensive lists of relations or it can be used to verify the completeness of present relations.

In addition to such a global list, it is also possible to apply specific strategies for extracting more comprehensive lists of word meanings related in a specific way. The most important relation in this respect is synonymy. In some cases these synonyms are explicitly listed in dictionaries but these specifications are not always complete or comprehensive. Several techniques are available for finding more candidates for synonymy:

- expanding from WordNet1.5.
- word meanings with similar definitions; one-word-definitions; circular definitions.
- overlapping translations in bilingual dictionaries.

The first technique is rather obvious. By directly translating the synset members in WordNet1.5 it is possible to derive synsets in another language. The second technique looks at definitions that are the very similar, and, in particular, definitions consisting of a single word or circularly defining words in terms of each other. This is illustrated by the following Dutch examples:

Table I. Association rate for *vino* (wine) in Spanish Dictionary

Association rate	***Frequency in dictionary***	***Paired word***
11.1655	15	*tinto* (red)
10.0162	23	*beber* (to drink)
9.6627	14	*mosto* (must)
8.6633	9	*jerez* (sherry)
8.1051	9	*cubas* (cask, barrel)
8.0551	16	*licor* (liquor)
7.2127	17	*bebida* (drink)
6.9338	12	*uva* (grape)
6.8436	9	*trago* (drink, swig)
6.6221	12	*sabor* (taste)
6.4506	15	*pan* (bread)

(13.) *apparaat* min of meer samengesteld **werktuig**
(apparatus) (more or less assembled tool)
instrument min of meer samengesteld of fijn **gereedschap** of **toestel** . . .
(instrument) (more or less assembled or delicate tool or apparatus)
toestel **apparaat**
(apparatus) (apparatus)
werktuig stuk **gereedschap**
(tool) (piece of tools)
gereedschap **werktuig**
(tools, instruments) (tool)

Here we see 5 different meanings that are circularly defined, suggesting a synonymy relation.

Another possibility is to look for words that have the same translations and/or occur as translations for the same words in bilingual dictionaries. The procedure is more or less as follows. Starting with a set of closely related Dutch words extracted on the basis of other techniques, such as the previous instrument examples *apparaat* (apparatus), *toestel* (apparatus), and *werktuig* (tool), and *gereedschap*(tools), we extract all the English translations for all their meanings from the bilingual Dutch-English dictionary. Next all these English translations are looked up in the reverse English-Dutch dictionary to see what Dutch words are given as translations for all the different meanings. The result is a very large list of translation-sets, covering very different meanings. However, we keep only those sets of Dutch translations that include at least two of the original words with which the search was started. These sets form a so-called translation-cycle via two bilingual resources. The co-occurrence of pairs of source words is thus used as a filter to select the correct meaning of the word. The automatically-generated result for the above words is the following list:

(14.) **Potential Equivalents generated from bilingual dictionaries:**

gebruiksvoorwerp 1	(implement, appliance, utensil)
comfort 1	(comfort)
mechanisme 2	(mechanism)
inrichting 5	(construction, installation)
tuig 1	(gear, equipment)
uitmonstering 3	(equipment, outfit, kit)
uitrusting 1	(equipment)
outillage 1	(equipment)
apparatuur 1	(apparatus, machinery)
materieel 1	(material, equipment)
machinerie 1	(machinery)
systeem 10	(system)
mechaniek 1	(mechanism)

Among them are a few synonyms but also words that can be related in other ways. (Atserias et al., 1997) describe a similar method for generating Spanish synsets.

Each of these techniques gives different results and requires further manual processing to achieve a coherent integration of the output. For example, the main source of data for the Italian wordnet is a combination of data from monolingual machine dictionary synonym fields and from a synonym dictionary, integrated with data from monolingual synonym-type definitions, and the semantic indicators in a bilingual Italian/English Lexical Data Base. All the data are extracted automatically but must be revised manually. Very briefly (and simplifying), the procedure for constructing the Italian synsets mainly operates in 3 steps:

1. Explicitly tagged synonyms contained in the machine-readable dictionary entries and synonym dictionary are grouped to form a first proposal of a synset. The output is revised manually.
2. Candidate synonyms extracted from synonym-type definitions (one-word definitions, similar definitions) are associated with all members of the synset under construction. The output is revised manually.
3. Each candidate for the synonym set is searched in the bilingual dictionary: semantic indicators and translation equivalents are associated and matched

against each other. The output is revised manually. A useful test for deciding whether a candidate belongs to a given synset is to examine the translation equivalent. If the translation equivalent for the doubtful item is very different from the translations of the other items in the synset, then it is likely that this item does not belong to the synset under construction.

The manual revision at the end of each stage is essential (see Roventini et al., 1998). After establishing a reasonable set of synonyms, the next problem is to find the relevant set of hyponyms. A selection of all words with the same genus word from a definition does not necessarily result in a coherent and comprehensive class. Due to alternative ways of defining or classifying meanings, words are spread over the hierarchies. The following main variations tend to occur (Vossen, 1995):

- Similar words are classified at different levels of abstraction.
- Different but more-or-less equivalent words have been used to classify the same meanings.
- Other perspectives have been chosen to classify similar meanings.

The first two variations are illustrated by the following examples from the Italian subset:

(15.)	*forchetta* (fork)	HAS_HYPERONYM	*arnese* (tool)
	coltello (knife)	HAS_HYPERONYM	*strumento* (tool)
	cucchiaino (teaspoon)	HAS_HYPERONYM	*posata* (piece of cutlery)

Here *cucchiaino* (spoon) is classified at an intermediate level as *posata* (piece of cutlery) which is then linked to the nearly equivalent classes *arnese* (tool) and *strumento* (tool), where you would expect to find all types of cutlery at the same level. The next example shows a variation in perspective:

(16.)	*avvelenare*	HAS_HYPERONYM	*uccidere*
	(to kill by poisoning)		(to kill)
	lapidare (to stone)	HAS_HYPERONYM	*colpire*
	(to kill by stoning)		(to hit)

Here we see that *avvelenare* (to poison) and *lapidare* (to stone) are classified within different hierarchies. This is the result of the way in which they have been defined in the monolingual dictionaries. Whereas *avvelenare* (to poison) is defined as “uccidere con il veleno” (to kill by means of poison), *lapidare* (to stone) is defined as “colpire con sassate per uccidere” (to hit with stones in order to kill). In both cases the result and the manner of achieving this are relevant but the Italian resources describe the events from different perspectives.

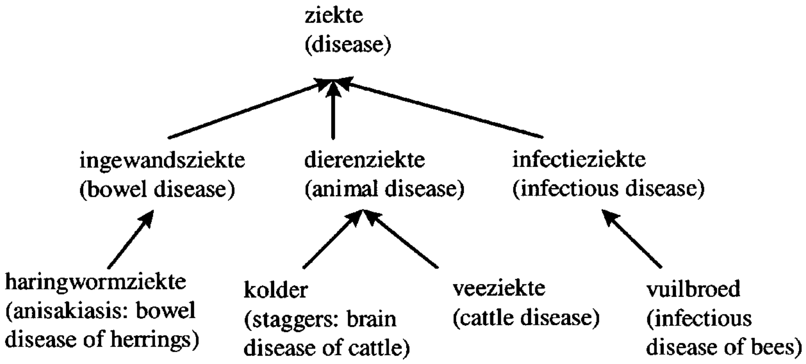

Figure 1. Hierarchical relations in the Van Dale database.

In the next hierarchy (Figure 1) containing Dutch words for diseases we see a typical combination of the phenomena, where multiple perspectives and levels have been missed. We see here that *haringwormziekte* (anisakiasis) is only linked to *ingewandziekte* (bowel disease) and *vuilbroed* (infectious disease of bees) is only linked to *infectieziekte* (infectious disease), while both are diseases of animals: *herrings* and *bees* respectively. In both cases, the classification as *dierenziekte* (animal disease) has been omitted. Within the same part of the hierarchy we see the opposite situation for *kolder* (staggers) which is directly linked to *dierenziekte* (animal disease) while it is also a disease of *cattle* and should be linked to *veeziekte* (cattle disease).

The hierarchy of diseases contains some typical examples of restructuring that are required because sub-levels of hyperonyms have been skipped and multiple classifications have been missed. Such variation in levels and multiple classifications can be detected by applying the Principle-of-Economy to the hyponyms (Dik, 1978). This principle states that it is not allowed to relate a word W1 to a word W3 when there is a word W2 linked to W3 to which W1 can be linked in the same way. In practice this means that all hyponyms of *ziekte* (disease) have to be cross-checked to see whether they represent hyperonyms of each other.[4] This then also reveals multiple category membership. When applied to the above cases we obtain the restructuring in Figure 2. Extracting information from different resources or merging different classification schemes gives a more comprehensive picture of a lexical semantic field but it also causes another problem. In some cases, the information given for these meanings is not coherent or exchangeable. This may either follow from the definitions of words which are supposed to be synonymous or be caused by the fact that the hyperonyms and/or hyponyms (or other semantic relations if present) do not apply to all the members of a synset. To some extent these problems are being tackled by individual measures such as the more-systematic encoding of multiple hyperonyms and the use of the so-called NEAR_SYNONYM relation.

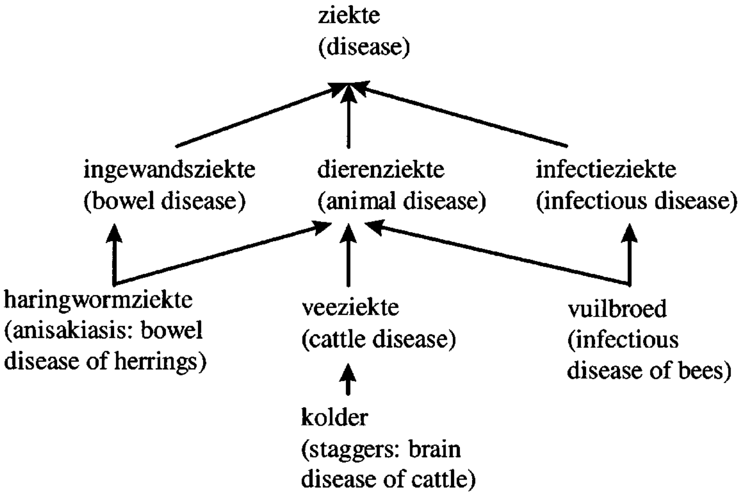

Figure 2. Restructured hierarchical relations.

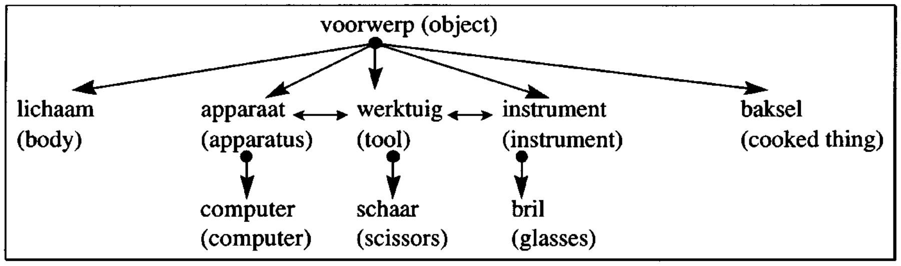

Figure 3. Near-synonymy relations between co-hyponyms.

In the case of the above example, where the Dutch words *apparaat* (apparatus), *toestel* (apparatus), *werktuig* (tool), and *gereedschap* (tools) have very similar and circular definitions, we may consider grouping them into a single synset. However, intuitively, they are not completely interchangeable, as is shown by the different clusters of hyponyms linked to them. Electrical devices are mostly classified as *apparaat* (apparatus), *instrument* (instrument), possibly as *toestel* (apparatus) but not as *werktuig* (tool) or *gereedschap* (tool). Instead of joining such closely-related meanings in a synset they can be related as NEAR_SYNONYMs so that they are distinguished from other co-hyponyms at the same level which are clearly not equivalent, while at the same time their hyponyms can be kept apart when they form different clusters. In Figure 3, we see that *apparaat* (apparatus), *werktuig* (tool) and instrument (*instrument*) still represent different clusters of hyponyms. The NEAR_SYNONYM relation expresses closeness, as opposed to other very different co-hyponyms like *baksel* (cooked thing) and *lichaam* (body).

2.3. OVERLAPPING RELATIONS

We have found that the tests do not always discriminate between all relations. This first of all shows itself in the subtypes of relations. As explained in (Alonge et al., this volume) the meronymy and role relations are differentiated into more general relations and more specific subtypes, such as HAS_MERO_MEMBER, has MERO_PORTION or HAS_ROLE_AGENT, HAS_ROLE_INSTRUMENT, etc. The more general relations are used when the more specific subtypes cannot clearly be assigned. Unclear cases of meronymy are the following examples:

(17.)	a	vlam 1	Portion?	*vuur* 2
		(flame)		(fire)
	b	*bloedfactor* 1	Made of?	*bloed* 1
		(blood factor)		(blood)
	c	*wijkgebouw* 1	Location?	*wijkcentrum* 1
		(building of community centre)		(community centre)

Portions normally are quantities of substances, e.g. a beer, two coffees, a snack. In the case of *vlam 1* (flame) and *vuur 2* (fire) it is however not clear whether we are dealing with a substance or with an event and hence it is unclear whether the meronymy relation portion can apply. In the case of (17)b it is not clear whether *bloedfactor 1* (blood factor) is a genuine component or a property, and a *wijkgebouw* 1 is both located at a *wijkcentrum* 1 and it is a part of it as well (they could even be synonymous). In such non-prototypical cases, where there is doubt about the specific relation, the most general relation HAS_MERONYM and HAS_HOLONYM is used.

As described in (Alonge et al., this volume), EuroWordNet distinguishes different roles or involvements of first-order-entities (concrete things) indicating arguments 'incorporated', or word meanings strongly implied, within the meaning of high-order entities (events). Most of these relations are (semi-)automatically extracted from regular definition patterns, such as "used for", "which causes", "a person who", "a place where", "made for", etc. However, we find examples where the extracted semantic roles are not prototypical, e.g.:

(18.)	a	*aardappelmoeheid*	Force/Cause	*aaltje*
		(potato disease)		(eelworms)
		antracose	Force/Cause	*steenkool*
		(anthracosis, miner's lungs)		(coal)
		betrekken	Force/Cause	*bewolking*
		(to cloud over)		(clouds)

	storen	Force/Cause	*hinder*
	(to disturb)		(disturbance)
b	*baarmoederhalskanker*	Location/Patient	*baarmoederhals*
	(cancer of the cervix)		(cervix)
	borstkanker	Location/Patient	*borst*
	(breast cancer)		(breast)
	bellenblazen	Patient/Result	*zeepbel*
	(blow bubbles)		(soap bubble)
	bespannen1	Patient/Result	*bespanning*
	(to string)		(stringing)
c	*verliezen2*	Agent/Patient	*verliezer*
	(to loose)		(a looser)
	winnen1	Agent/Patient	*winnaar*
	(to win)		(a winner)

In (18)a we see some examples where a concrete entity causes a situation but it cannot be seen as an Agent having any control or intention of doing this. However, since the causes relation is restricted to higher-order-entities (events, states) it cannot be applied here. The relation between e.g. *aaltjes* (eelworms) and the disease is in fact more indirect. The *eelworms* only create the circumstances, which result in the disease. The same holds for *clouds*, *coal* and *disturbance*, they are Factors, Forces or Causes but not Agents. Here we can either broaden the interpretation of Agent or add new roles. In (18)b we see cases where the Patient-role interferes with other roles. In the first two examples we see an entity with a double role as the affected entity (by a disease) but also as the location where the disease is active. They could be considered as Location or as Patient. Another group of dubious Patients are entities which are created by some event. As concrete entities, they cannot be related by means of a CAUSES relation but they can still be seen as the result of an event. Again we can choose to broaden the interpretation of Patient or add a new relation. Finally, in (18)c we see two typical examples where an entity is actively involved in a (competition) event, but has no control over the outcome and is conceptualised as the affected entity (positive or negative) as well. In this case, we can decide to allow both the Patient and Agent relation, although it is still not a prototypical Agent having control over the action. In all the above cases, we have now decided to use the under-specified relation ROLE. The advantage of the under-specified relation is clear. The lexicographer does not have to solve a complex problem to continue with an isolated case, whereas all the undifferentiated relations can be collected at a later stage and regular patterns can be differentiated after reaching agreement with the other sites in the project.

More serious than under-differentiation of relations are cases where incompatible relations still show some overlap in interpretation. This is the case for two classes of relations: hyponymy/synonymy versus meronymy/subevent, and agent/instrument roles versus CAUSES. In the following examples we see meanings where one entity or event consists of components or subevents but is also hardly distinguishable from it:

(19.)	a	sports	HAS_SUBEVENT?	sport-game
	b	bevolkingsgroep	HAS_PART_MEMBER	bevolking
		(group of people)		(people)
		gebladerte	HAS_PART_MEMBER	blad
		(leafage)		(leaf)
		gesteente	HAS_PART	steen
		(stones)		(stone)

In (19)a we see a complex event or activity which consists of the subevent *sport-game*, but the difference is subtle. Especially when pluralized, a subevent can easily be used to replace the larger event that includes it. Differences in number are not reflected by one of the semantic relations in EuroWordNet. The same holds for the meronymy relation in (19)b. The group-noun *bevolkingsgroep* and the collective *bevolking*, as well as *gebladerte* and the plural form *bladeren* (leaves) are denotationally equivalent (can refer to the same type of entities), but differ in grammatical reference. In the case of the collective *gesteente* (stones, especially as a kind of stone) and the mass noun *steen* (stone), we see that the difference is only the genericity of reference. In all these cases, it is difficult to decide on synonymy/hyponymy on the one hand and meronymy/subevent on the other (Vossen and Copestake, 1993; Vossen, 1995). Because of the homogeneity of the composition we often see that both the complex concepts and the component are linked to the same hyperonym as well. For example, both *mood* and *feeling* are subtypes of mental state, and both *gesteente* (stones) and *steen* (stone) are linked to *stof* (substance).

When discussing the role/involved-relations we more or less suggested that there is a close relation between agents and instruments on the one hand and CAUSES-relations on the other hand. So far we have stated that the former relate first-order-entities to dynamic events, whereas the latter can only be used to relate high-order-entities. However, the distinction between first-order-entities and high-order-entities is not always clear-cut, and this results in cases where the difference between agent-roles and CAUSES starts to fade as well. There are three ways in which there can be lack of clarity about the status of an entity:

1. words may refer to properties and to concrete entities having that property.
2. non-concrete words such as thoughts, ideas, opinions still have entity-like properties.
3. words may vary over both types of entities.

We have discussed examples where some process or change results in a concrete entity and a similar change or process may also result in a state as in (20)a. However, in some cases the word naming the result refers to both the state or an entity in such a state, as in (20)b:

(20.)	a	*verwoestijnen*	INVOLVED	*woestijn*
		(to become a desert)		(desert)
		evaporar		*vapor*
		(to evaporate)		(vapour)
		natworden	CAUSES	*nat*
		(to become wet)		(wet)
		afear		*feo*
		(to make ugly)		(ugly)
	b	*mineralize*	CAUSES/INVOLVED	*mineral*
		liquidify	CAUSES/INVOLVED	*liquid*
		solidificar	CAUSES/INVOLVED	*sólido*
		(to become solid)		(solid)

Here we see that *mineral* and *liquid* can be both a noun and an adjective denoting a substance or a state of a substance and the intuitive interpretation does not differ much from both examples in (20)a. For such resultative events we have taken an arbitrary position that the classification of the result as first or high-order-entity is the only criterion: i.e. if *mineral* is disambiguated as a noun the relation will be INVOLVED, just as for *desert*, if it is an adjective the relation will be CAUSES, just as for *wet*.

A second problematic case is represented by words denoting sounds, mental states or objects which are not concrete first-order-entities but share a lot of properties with them:

(21.)	a	*musiceren1*	CAUSES/INVOLVED	*muziek*
		(make music)		(music)
		zingen1		*lied*
		(sing)		(song)
		cantar		*canción*
		(sing)		(song)
	b	*bekeren* 5	CAUSES/INVOLVED	*mening*
		(convert, reform)		(opinion)
		bedenken 1		*gedachte*
		(think up)		(thought)

	juzgar		*juicio*
	(to judge)		(judgement)
c	*nominaliseren*	CAUSES/INVOLVED	*naamwoord*
	(nominalize)		(noun)
	nominalizar		*nombre*

In (21)a we see that the relation for production of sound depends on how sounds are treated. In EuroWordNet they are classified as higher-order-entities so strictly speaking the relation should be CAUSES. However, if considered as a physical signal the same criterion would predict that there should be an INVOLVED relation. In (21)b we see that a mental or communicative event results in a mental state or thought and again the status of these as entities determines the type of relation. Metaphorically, thoughts and opinions are very much like concrete entities. You can work on them, create them, keep them, multiply them, etc. Therefore we have applied the relation INVOLVED here. Finally, (21)c represents a difficult case because the result is a word which can be a symbolic representation, a sound, or a concept in the mind, where the former is a first-order-entity and the latter two are high-order-entities.

Another example where ROLE and CAUSE relations converge is represented by words referring to the initiator of an event without implying further information on the status of the entity. For example, the Dutch noun *middel 1* (means) can stand for any event, method, or instrument leading to some change:

(22.)	*middel*1	INSTRUMENT/CAUSES	*veranderen*
	(any means or method to achieve something)		(to change, alter)

Clearly, the level-of-entity criterion does not work here. Related to this are so-called Modal-states which are properties or situations which are necessary conditions or qualities to make a change or event possible. Typical examples of these states are mental and physical abilities:

(23.)	*gehoor*	CAUSES	*horen*
	(hearing: the capability to hear)		(to hear)
	mogelijkheid	CAUSES	*gebeuren* 2
	(possibility)		(to take place)
	visión	CAUSES	*ver*
	(vision)		(to see)
	sentido		*sentir*
	(sense)		(to feel)

The relation between the capacity and the associated event is now expressed by means of a CAUSES relation in EuroWordNet.

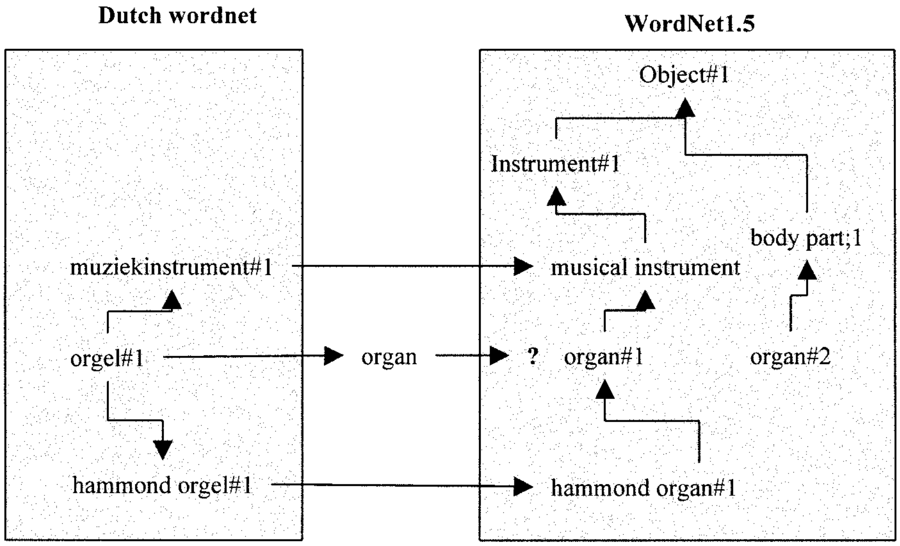

Figure 4. Selecting translations to WordNet1.5 by conceptual distance measuring to the translated context in the Dutch wordnet.

Concluding, we can say that the notion of causality applies to a wide range of relations, from genuine cause relations between events and results, to agents and instruments, to modal states or abilities. In between the clearer cases, there are many meaning relations which are not easy to classify.

3. Establishing Equivalence Relations

The second type of problems regard the specification of equivalence relations. As stated in the introduction of this volume, each synset in the monolingual wordnets will have at least one equivalence relation to a concept in the Inter-Lingual-Index (ILI). Especially, at the start of the project this ILI mainly consisted of synsets taken from WordNet1.5 synsets. The linking to WordNet1.5 is partly done using automatic techniques and partly manually. For example, the translations for most Spanish nouns are generated automatically on the basis of the following criteria:

- monosemous translations of synsets with a single sense are directly taken over as translations.
- polysemous translations are disambiguated by measuring the conceptual-distance in the WordNet1.5 between the senses of multiple translations (Agirre and Rigau, 1996; Rigau et al., 1997).

The latter technique calculates the distance between two senses by counting the steps to their closest shared node in the network, taking into account the level of the hierarchy and the density of nodes relative to the average density. When two translations are given for a Spanish word and these translations have multiple

Table II. Reliability of the automatically-generated equivalence relations in the Dutch wordnet

	Nouns		***Verbs***	
Matching Rank	No. of synsets	Percentage	No. of Synsets	Percentage
1st score	70	70.71%	20	40.82%
2nd score	14	14.14%	13	26.53%
3rd score	5	5.05%	9	18.37%
>	1	1.01%	3	6.12%
lexical gaps	7	7.07%	1	2.04%
no correct	20	20.20%	12	24.29%
Total of synsets	99[5]		58	

senses in WordNet1.5, those senses are selected which have the shortest distance in the hierarchy. A similar approach has been applied to the Dutch and Italian wordnet but in this case we took advantage of the translated context in the hierarchy as well (Vossen et al., forthcoming; Peters et al., forthcoming), as is illustrated in Figure 4 for Dutch. Here we see that *orgel* in Dutch is translated as *organ*, which can either be a musical instrument or a body part. Since the hyperonym and a hyponym of *orgel* in the Dutch wordnet have already been translated it is possible to measure the distance of the two senses of organ to the translations of the hyperonym and hyponym.

The distance measuring of the translations to the context in the Dutch wordnet, leads to a ranking of all the senses of a translation. Table II gives the reliability of this methodology for a random sample of 99 nominal and 49 verbal synsets. The score for each ranking indicates the number of synsets that are the correct translations. In most cases of the nouns (71), the highest translation is the correct one. In 20% of all nouns, the correct translation was not among the proposed translations at all. In 7% of all nouns, there was no good translation possible (lexical gaps), because the meaning does not exist in English or in WordNet1.5. For the verbs the results are considerably worse. Only 41% of the highest ranking was correct. This difference is the result of the fact that the verb-hierarchies are more shallow and diverse. If many verbs are linked to the same hyperonym or too many different but unrelated tops in WordNet1.5 this results in a poor matching for all candidates. Note, however, that by taking the top-3 ranking, the results for nouns and verbs are about the same (90% versus 85.6%). In the case of verbs, it appears to be difficult to choose and several senses of the translations could apply. It thus makes sense to select the best 3 translations for verbs instead of trying to select a single best sense.

As these figures show, a manual revision of suspect cases is necessary. Furthermore, crucial meanings are encoded manually in the first place. There are three main problems that play a role when establishing these equivalence relations which we will discuss in more detail below:

Table III. Matching of Spanish-English bilingual dictionary with WordNet1.5

	English nouns	Spanish nouns	synsets	connections[6]
WordNet1.5	87,642	—	60,557	107,424
Spanish/English	11,467	12,370	—	19,443
English/Spanish	10,739	10,549	—	16,324
Merged Bilingual	15,848	14,880	—	28,131
Maximum Reachable Coverage[7]	12,665	13,208	19,383	66,258
Of WordNet	14%	—	32%	—
Of bilingual	80%	90%	—	—

- lexical gaps;
- differences in sense-differentiation;
- fuzzy-matching;

These problems not only show up in the automatic matching of synsets to WordNet1.5 but also when we try to assign the equivalence relations manually.

3.1. LEXICAL GAPS

Gaps may either be due to inadequacy of the resources or to differences in lexicalization across the languages. Four specific problems may occur (Copestake et al., 1995):

- there may be no entry
- there may be a phrasal translation in a bilingual dictionary (phrases, compounds, derivations, inflected forms).
- the translation is not an entry in WN1.5,
- the intended sense of a translation is not present in WN1.5 (although the word itself is).

We will illustrate these problems for the Spanish lexical resources (see Atserias et al., 1997, for further details). By merging both directions of the nominal part of the Spanish/English bilingual dictionaries we obtained an homogeneous bilingual dictionary (that is, both directions of a bilingual dictionary are normally not symmetric). As is shown in the Table III, the maximum coverage we can expect using this small bilingual dictionary, ranges from 14% of all WN1.5 nouns to 32% of WN1.5 synsets (including errors). On the other hand, this mapping does not yield a connection to WN1.5 for 20% of the English nouns appearing in the homogeneous dictionary and 10% of the Spanish words.

The simplest mapping presented in (Atserias et al., 1997) is the situation where a Spanish word has a unique English translation in both directions and this English

Table IV. Overlap in lexical units across monolingual and bilingual sources

A	Noun Definitions	93,394	
B	Noun Definitions with Genus Word	92,693	
C	Genus Words	14,131	
D	Genus with Bilingual translation	7,610	54% of c)
E	Genus with WordNet translation	7,319	52% of c)
F	Headwords	53,455	
G	Headwords with Bilingual translations	11,407	21% of f)
H	Headwords with WordNet translations	10,667	20% of f)
I	Definitions with Bilingual translations	30,446	33% of b)
J	Definitions with WordNet translations	28,995	31% of b)

word has only one sense in WN1.5. Only 92% of the connections produced by this method were considered correct. Another 2% of the connections were considered hyponyms of the correct ones, 2% nearly correct and 2% fully incorrect. Examples of correct and incorrect connections are the following. For instance *horn* could be translated in Spanish as *asta*, *bocina*, *claxon*, *cuerno*, etc. *Horn* in Spanish has (at least) two meanings: part of an animal and part of a car. As the homogeneous bilingual dictionary only connects words (not meanings) the following connections could be produced.

00740047 05 *horn asta* OK of an animal
00740047 05 *horn bocina* ERROR of an animal (OK of a car)
00740047 05 *horn claxon* ERROR of an animal (OK of a car)
00740047 05 *horn cuerno* OK of an animal

Another problem relates to differences in size of monolingual and bilingual resources that are merged. Table IV shows the overlapping across lexical units and resources. The monolingual dictionary contains 93,394 noun definitions (a), relating 53,455 headwords (f) and 14,131 genus words (c). Whereas there is a bilingual translation for 54% of the genus words, the bilingual dictionary only covers 21% of the headwords. The mapping only produces fully connected definitions (both headword and genus word) for 33% of the whole monolingual source. Furthermore, approximately 2% of the Spanish lexical units cannot be mapped to WN1.5 because the English translation was not found.

If there is no translation or only a phrasal translation for a sense in the dictionary it may be the case that we are dealing with a lexical gap. There may be different types of lexical gaps:

- Cultural gaps, e.g. the Dutch verb: *klunen* (to walk on skates) refers to an event not known in the English culture.
- ‘Pragmatic gaps’, e.g. the Dutch compound verb form *doodschoppen* (to kick to death), the Spanish *alevín* (young fish), or the Italian verb *rincasare* (to go

back home), which all refer to concepts known in the English culture but not expressed by a single lexicalized form. In these cases the lexicalization patterns in the languages are different from English.

- Morphologic mismatches: e.g. in Dutch the adjective *aardig* is equivalent to the verb *to like* in English.

In all these cases the Source Synset is linked to the closest Target-equivalent using a so-called complex-equivalence relation. Complex-equivalence relations parallel the language-internal relations (HAS_EQ_HYPERONYM, HAS_EQ_MERONYM, etc.). In most cases a lexical gap will be related to a more general concept with a HAS_EQ_HYPERONYM relation. In the case of the morphological gap EuroWordNet provides the possibility to encode a cross-part-of-speech equivalence relation. Likewise there can still be an EQ_SYNONYM relation between *aardig* Adjective and *like* Verb:

(24.) Equivalence relations for Gaps

Dutch WordNet	**Equivalence Relation**	**WordNet1.5**
klunen (to walk on skates)	HAS_EQ_HYPERONYM	*Walk*
aardig	EQ_SYNONYM	*Like*

3.2. SENSE-DIFFERENTIATION ACROSS WORDNETS

The second problem is that matching entries across resources shows differences in the differentiation of senses. Obviously, this problem is related to the sense-differentiation problems discussed above. Again we can make a distinction between under-differentiation and over-differentiation, which can occur either at the source wordnet or the target wordnet (in the case of EuroWordNet synsets taken from WordNet1.5):

Over-differentiation

- multiple targets: Dutch *schoonmaken* has only 1 sense whereas English *clean* has 19 senses. Here WN1.5 gives senses for different pragmatic uses that should not be distinguished as separate senses. The target is clearly over-differentiated.
- multiple sources: Dutch *versiersel* and *versiering* are both linked to the same WN1.5 synset *decoration* but are still distinguished as different synsets in the Dutch resource. There is however no difference in their definition or any other information. Here the source is over-differentiated.

Under-differentiation

- multiple targets: The Dutch sense *keuze* is defined as the *act* or *result* of choosing, likewise it can be linked both to *choice* 1 the act of choosing and *choice* 2 what is chosen. Two incompatible Dutch senses are conflated: the source is under-differentiated.

- multiple sources: *hout 1* (wood as substance), *houtsoort 1* (kind of wood) / wood 4. WN1.5 gives only one sense for wood, which has to capture both meanings *kinds* of wood and a *portion*. The target is under-differentiated (although it is less clear whether this is a mistake).

To solve these matching problems we are taking some specific measures. First of all the EQ_SYNONYM relation is only used when there is a clear and simple equivalence relation with a single synset in another resource (either at the source-side or the target-side). When there is no partial overlap or matching with a target synset, the source-synset is treated as a lexical gap in WordNet1.5 until we find evidence to the contrary. In the case of too many and too fine-grained sense-distinctions in the target or source-wordnet we agreed to apply the EQ_NEAR_SYNONYM relation. This would apply to the above case where a single sense in Dutch matches multiple senses of *clean*:

(25.) Near-Equivalence relations to multiple targets

Dutch WordNet	**Equivalence Relation**	**WordNet1.5**
schoonmaken 1	EQ_NEAR_SYNONYM	Clean 1 (making clean by removing filth, or unwanted substances)
shoonmaken 1	EQ_NEAR_SYNONYM	Clean 2 (remove unwanted substances from, such as feathers or pits, as of chickens or fruit)
schoonmaken 1	EQ_NEAR_SYNONYM	Clean 7 (remove in making clean)
schoonmaken 1	EQ_NEAR_SYNONYM	clean 8 (remove unwanted substances from – (as in chemistry))
hout 1 (wood as substance)	EQ_NEAR_SYNONYM	wood 4.
houtsoort 1 (kind of wood)	EQ_NEAR_SYNONYM	wood 4.

Obviously, judging the differences in sense-differentiation as over-differentiation or under-differentiation will eventually lead to a restructuring of the sense-differentiation of the source-wordnets and WordNet1.5. The cases of under-differentiation have in fact already been discussed in the previous section. Whenever conflated hyperonyms are incompatible (e.g. according to a co-ordination test) the sense will have to be split into two separate senses. In the case of over-differentiation we will see to what extent it is possible to globalise the sense-differentiation. In the case of WordNet1.5 this is particularly important because over-differentiation may cause equivalent meanings across wordnets to be linked to different WordNet1.5 senses.

Another sense-differentiation problem has again to do with the inconsistent treatment of regular polysemy across resources. In the next examples we see that the Dutch resource lists two senses for both *ambassade* (embassy) and *academie* (academy), one as the building and one as the institute, while WordNet1.5 specifies only one sense for each, but a different one:

(26.)

NL-WordNet		**WordNet1.5**
ambassade	*1 <organization>*	0
	2 <building>	*embassy*
academie	*1 <organization>*	academy
	2 <building>	0

These regular patterns of polysemy can also be generated to partially overcome the inconsistent listing of senses across resources. This solution has been applied by (Hamp and Feldweg, 1997) in the building of the German wordnet, by encoding a polysemy-relation between classes of concepts that exhibit regular meaning-shifts (animal-food, institute-building, animal-human, etc.). The advantage is not only that omissions may be corrected but also that mismatchings across resources may be resolved. If for example the Dutch resource represents *universiteit* (university) as the institute and the Spanish resource represent *universidad* as the building, the regular polysemy pointer will generate the missing senses for both resources:

(27.) Metonymic Equivalence relations

Dutch WordNet	**WordNet1.5 Equivalents**	**Spanish WordNet**
universiteit	University the institute	<extended meaning>
<extended meaning>	University the building	*universidad*

In EuroWordNet we will extend the ILI with global synsets that represent groups of senses related either as specializations of a more general meaning or by means of regular polysemy as above. In (Peters et al., this volume) we discuss in detail how specific synsets in the wordnets can be related to these more global synsets.

3.3. MISMATCHES OF SENSES

A final case of mismatching to be discussed is the situation in which there is a close match with a specific target synset but the information across the wordnets does not match. The mismatching information could be:

- the way the meanings are classified (their hyperonyms are not equivalent or different hyponyms are listed), e.g.:

(28.)	a	**NL-WordNet**		
		hond	HYPERONYM	*huisdier*
		(dog)		(pet)
		WordNet1.5		
		dog	HYPERONYM	*canine*
	b	**SP-WordNet**		
		queso	HYPERONYM	*masa*
		(cheese)		(substance)
		WordNet1.5		
		cheese	HYPERONYM	dairy_product

Here the mismatching depends on the compatibility of the hyperonyms (see discussion above). Only when the hyperonyms cannot be combined as conjuncted predicates may it be necessary to reconsider the equivalence relation. In these examples both classifications are acceptable (*a dog is a pet and a canine*; *cheese is a mass and a dairy product*).

Obviously, differences in classification also lead to situations in which two equivalent hyperonyms have different sets of hyponyms below them. In the above cases we can expect that Dutch *huisdier* and English *pet*, or Spanish *masa* and English *substance* will differ in the hyponyms but may still have equivalent definitions and hyperonyms themselves. The differences in these examples do not falsify the equivalence relations but only show that the classifications differ (either as an inconsistency or as a language-specific property).

- their definitions may deviate in some way;

(29.)	**IT-WordNet**	**WordNet1.5**
	seguace, descepolo	*follower*
	(=who strongly believes)	*(=who accepts)*

Here we see that the gloss for the Italian synset is more specific than the English gloss, despite the equivalence relation in bilingual dictionaries. This difference may still fall within the limits of acceptable variation and the equivalence relation is legitimate.

- they may differ in the synset-members;

This is very likely to happen when large synsets are mapped. Comparison of both wordnets shows that in many cases there are large synsets in both languages for the same concepts, but these often are not parallel. Differences are mostly due to unbalanced differentiation in both wordnets. For example *onzin 1* (nonsense) in the Dutch wordnet has 36 synset members, possible candidates as equivalents in WN1.5 are *humbug 1* (10 synset members) and *bullshit 1* (13 synset members). These are however represented in different synsets in WordNet1.5:

(30.) <subject matter1, message2, content3, substance4>
    what a communication that is about something is about
    HAS_HYPONYM:
    —
    <nonsense2, meaningless2, nonsensicality1>
        HAS_HYPONYM
        <humbug1, baloney1, bilgewater1, boloney1, bosh1, drool2,
        tarradiddle1, tommyrot1, tosh1, twaddle1>
    <drivel2>
    a worthless message
        HAS_HYPONYM
        <Irish bull1, bull3, *bullshit1*, buncombe1, bunk2, bunkum1,
        crap1, dogshit1, guff1, hogwash1, horseshit1, rot1, shit3>

To distinguish *bullshit 1* as a worthless message from *baloney 1* as *nonsense 2* looks like over-differentiation of WordNet1.5. In the Dutch wordnet however the synset of *onzin 1* is extremely large. It contains words like *gekakel 2* (cackle/chatter), *gezwam* (empty talk), which are not synonyms of *onzin* (nonsense) but more specific hyponyms. So here there is under-differentiation as well at the Dutch side.

Obviously, in all the above cases there must be something in common to seriously consider an equivalence relation. In general we follow the policy that we take the concept or gloss as the starting point. Differences in hyperonyms or hyponyms can also be caused by other reasons. To indicate a less precise matching these synsets should always be linked with an EQ_NEAR_SYNONYM relation.

4. Conclusions

In this paper we have described a general procedure for building wordnets in EuroWordNet, discussing the major problems that may be encountered, especially when dealing with the more complicated Base Concepts. The decisions taken for these words have an effect on the structure of the database as a whole. By following a common strategy and shared solutions we ensure that these fundamental building blocks are encoded in a similar way across the different wordnets.

Usually, a summary of problematic examples is a disappointing enterprise. However, it is important to realise that not all meanings and relations are as complicated as suggested here. In many cases the relations are obvious and most words only have one or two meanings. Large fragments of the wordnets are therefore generated (semi-)automatically looking for patterns in definitions, mapping synsets via bilingual dictionaries or comparing taxonomies. These procedures are not discussed here, but will be described in a separate deliverable of the project on the tools and methods for building the wordnets.

Notes

[1] Note that it is not allowed to list two senses of the same entry in the same synset. Two senses can therefore only be merged in a variant of the same synset by deleting one sense and adding the related information to the remaining sense.

[2] We often see that disjunctive hyperonyms (hyperonyms that cannot apply simultaneously) form a regular metonymic pattern or alternation pattern. In principle their senses should be separated although it is possible to keep the collapsed meaning as well. In Peters et al. (this volume) we will discuss how these regular polysemy patterns can be captured via collapsed synsets in the Inter-Lingual-Index, regardless of the way they are treated in the individual wordnets.

[3] Association Ratio can be defined as the product of Mutual Information by the frequency. Given two words w1 and w2 which co-occurs in some definitions:

> AR(w1,w2) = Pr(w1,w2)∗log(Pr(w1,w2)/Pr(w1)∗Pr(w2)) where Pr(w1,w2) is the estimation of the probability of w1 and w2 co-occur in some definitions and Pr(w) is the estimation of the probability of w occur in some definition.

[4] Some practical strategies for finding similar meanings which are classified differently, is by making use of the morphology of the entries (e.g. compounds ending with *disease*), or by looking for other, alternative definition patterns (e.g. containing phrases such as *infectious*).

[5] The total of scores exceeds the total of synsets because in some cases multiple senses or translations appear to be correct.

[6] Connections can be word/word or word/synset. When there are synsets involved the connections are Spanish-word/synset, (except for WordNet itself), otherwise Spanish-word/English-word.

[7] Maximum Reachable Coverage. Given the translations placed in the bilingual we can only attach Spanish words to 32% of WN1.5 synsets, 14% of WN1.5 nouns, etc. This is the maximum we can reach (most of these connections could be wrong).

References

Antelmi D. and A. Roventini. “Semantic Relationships within a Set of Verbal Entries in the Italian Lexical Database”. In *Proceedings of EURALEX ’90*, IV International Congress, Benalmadena (Malaga), 28–8/1–9, 1990.

Agirre E. and G. Rigau. “Word Sense Disambiguation using Conceptual Density”. In *Proceedings of the 16th International Conference on Computational Linguistics* (COLING’96). Copenhagen, Denmark, 1996.

Atkins B. and B. Levin. “Admiting Impediments”. In *Proceedings of the 4th Annual Conference of the UW Centre for the New OED*, Waterloo, Canada, 1988.

Atserias J., S. Climent, X. Farreres, G. Rigau and H. Rodríguez. “Combining Multiple Methods for the Automatic Construction of Multilingual WordNets”. In *Proceedings of International Conference “Recent Advances in Natural Language Processing”*, Tzigov Chark, Bulgaria, 1997.

Cruse, D. A. *Lexical Semantics*. Cambridge: CUP, 1986.

Dik, S. *Stepwise Lexical Decomposition*. Lisse: Peter de Ridder Press, 1978.

Gale W., K. Church and D. Yarowsky. “A Method for Disambiguating Word Senses in a Large Corpus”. *Computers and the Humanities*, 26 (1993), 415–439.

Hamp, B. and H. Feldweg, “GermaNet: a Lexical-Semantic Net for German”. In *Proceedings of the ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Eds. P. Vossen, N. Calzolari, Adriaens, Sanfilippo and Y. Wilks, Madrid, 1997.

Jacobs, P. “Making Sense of Lexical Acquisition”. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, Ed. Zernik U., Hillsdale, New Jersey: Lawrence Erlbaum Associates, publishers, 1991.

Levin, B. *English Verb Classes and Alternations*. Chicago: University of Chicago Press, 1993.

Peters, C., A. Roventini, E. Marinai and N. Calzolari. “Making the Right Connections: Mapping between Italian and English Lexical Data in EuroWordNet”. In *Proceedings of the Joint International Conference ALLC/ACH ’98* “Virtual Communities”, 5–10 July 1998, Lajos Kossuth University, Debrecen, Hungary (forthcoming).

Rigau G., J. Atserias and E. Agirre. “Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation”. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (ACL’97). Spain: Madrid, 1997, pp. 48–55.

Roventini, A. “Acquiring and Representing Semantic Information from Place Taxonomies”. *Acta Linguistica Hungarica*, 41(1–4) (1992), 265–275.

Roventini, A., F. Bertagna, N. Calzolari and C. Peters. “Building the Italian component of EuroWordNet: A Language-specific Perspective”. *Proceedings of Euralex ’98*, August, Brussels, Belgium (forthcoming).

Vossen P. and A. Copestake. “Untangling Definition Structure into Knowledge Representation”. *Default Inheritance in Unification Based Approaches to the Lexicon*. Eds. E.J. Briscoe, A. Copestake and V. de Paiva. Cambridge: Cambridge University Press, 1993.

Vossen P. *Grammatical and Conceptual Individuation in the Lexicon*, PhD. Thesis University of Amsterdam, Studies in Language and Language Use, No. 15. IFOTT, Amsterdam, 1995.

Vossen, P., L. Bloksma and P. Boersma. “Generating Equivalence Relations to WordNet1.5 by Aligning the Hierarchical Context”. In *Proceedings of the Workshop on Cross-language Semantic Links*, organized by the Institut fuer Deutsche Sprache, Pescia, 19th–21st June 1998 (forthcoming).

Wilks Y., D. Fass, C. Guo, J. McDonal, T. Plate and B. Slator. “Providing Machine Tractable Dictionary Tools”. *Semantics and the Lexicon*. Ed. J. Pustejowsky, Dordrecht: Kluwer Academic Publishers, 1993, pp. 341–401.

Zwicky A. and J. Sadock. “Ambiguity Tests and How to Fail Them”. *Syntax and Semantics 4*. Ed. J. Kimball, New York: Academic Press, 1975.